Information Sciences 278 (2014) 327–343

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Expressive generalized itemsets

Elena Baralis, Luca Cagliero, Tania Cerquitelli, Vincenzo D'Elia, Paolo Garza *

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

Article history:
Received 15 April 2013
Received in revised form 20 December 2013
Accepted 11 March 2014
Available online 24 March 2014

Keywords:
Generalized itemset mining
Data Mining
Expressiveness of generalized itemset

ABSTRACT

Generalized itemset mining is a powerful tool to discover multiple-level correlations among the analyzed data. A taxonomy is used to aggregate data items into higher-level concepts and to discover frequent recurrences among data items at different granularity levels. However, since traditional high-level itemsets may also represent the knowledge covered by their lower-level frequent descendant itemsets, the expressiveness of high-level itemsets can be rather limited. To overcome this issue, this article proposes two novel itemset types, called Expressive Generalized Itemset (EGI) and Maximal Expressive Generalized Itemset (Max-EGI), in which the frequency of occurrence of a high-level itemset is evaluated only on the portion of data not yet covered by any of its frequent descendants. Specifically, EGI s represent, at a high level of abstraction, the knowledge associated with sets of infrequent itemsets, while Max-EGIs compactly represent all the infrequent descendants of a generalized itemset. Furthermore, we also propose an algorithm to discover Max-EGIs at the top of the traditionally mined itemsets.

Experiments, performed on both real and synthetic datasets, demonstrate the effectiveness, efficiency, and scalability of the proposed approach.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Frequent itemset extraction [1] is a widely used exploratory technique to discover interesting correlations among data items. A frequent itemset is a set of data items, whose observed frequency of occurrence (support) in the source dataset is above a given threshold. A taxonomy (i.e., a set of is-a hierarchies) built over the data items can be used to aggregate items, based on granularity concepts, into higher-level items, called generalized items. Taxonomies enable the discovery of multiple-level patterns from the analyzed data. This process is called frequent generalized itemset mining [27]. Generalized itemsets are itemsets that may contain either data items or generalized items defined in the taxonomy. In most related works (e.g., [6,19,26]), the analyzed datasets are equipped with user-provided taxonomies and the corresponding generalized items are defined according to domain-specific knowledge. For example, in the healthcare domain examinations and drugs can be generalized as the corresponding categories, while in market basket analysis products (items) can be aggregated into the corresponding category or brand [27]. Frequent generalized itemsets are generalized itemsets whose support is above a given threshold. The support is defined as the ratio between the number of dataset records covered by the itemset and the total number of records in the dataset. The knowledge represented by a high-level (generalized) itemset is the same as a set of low-level descendants, including frequent and infrequent itemsets. However, a frequent high-level itemset is extracted even if its corresponding subset of frequent low-level itemsets covers almost the same dataset records. Hence, there is a need for improving the expressiveness of traditional generalized itemsets by proposing new types of multiple-level patterns.

* Corresponding author. Tel.: +39 011 090 7022; fax: +39 011 090 7099.
E-mail addresses: elena.baralis@polito.it (E. Baralis), luca.cagliero@polito.it (L. Cagliero), tania.cerquitelli@polito.it (T. Cerquitelli), vincenzo.delia@polito.it (V. D'Elia), paolo.garza@polito.it (P. Garza).

http://dx.doi.org/10.1016/j.ins.2014.03.056
0020-0255/© 2014 Elsevier Inc. All rights reserved.

Table 1
Example dataset.

Order date	Clothing shop	City	Most discounted item
2012-06-17	Shop_A	Turin	Jackets
2012-09-01	Shop_A	Turin	Hiking boots
2012-10-20	Shop_B	Cambridge	Ski pants
2012-05-01	Shop_C	Rome	Hiking boots
2012-05-01	Shop_C	Tivoli	Jackets
2012-10-20	Shop_B	Asti	Hiking boots
2012-10-20	Shop_A	Cuneo	Jackets
2012-05-01	Shop_A	Cuneo	Hiking boots
2012-10-20	Shop_C	Alba	Gloves

This article presents (i) two new itemset types, namely the Expressive Generalized Itemset (denoted by EGI) and the Maximal Expressive Generalized Itemset (Max-EGI). Both EGI and Max-EGI extend the notion of generalized itemset [27] by enhancing the pattern expressiveness with respect to its descendant set. (ii) An algorithm to discover Max-EGI s at the top of traditional itemsets. (iii) An in-depth experimental evaluation on many structured datasets to demonstrate the effectiveness and efficiency of the proposed approach to mine interesting and highly expressive patterns. (iv) An example of application of the proposed approach to a real-life application context, i.e., the analysis of network traffic captures. The proposed approach can be profitably exploited to analyze data coming from different application contexts if real data can be equipped with meaningful taxonomies (e.g., context-aware applications, network traffic data analysis, medical treatments).

As an example, Table 1 reports a running example dataset, where each record stores some information about the orders submitted to a chain of clothing shops. For each order the date, the name of the shop who received the order, the city from which the order has been submitted, and the most discounted item are stored. Fig. 1 shows a simple taxonomy built on the set of attributes of the running example. For the sake of clarity, in this section we consider only the subset of frequent generalized itemsets mined from the *most discounted item* attribute of the example dataset. Table 2(a) reports the set of frequent generalized itemsets mined by a traditional mining algorithm [27] by enforcing an absolute minimum support threshold equal to 2 and by exploiting the taxonomy in Fig. 1. Both the "low-level" itemset (e.g., {(Most discounted item,Hiking boots)}) and the "high-level" (generalized) itemset (e.g., {(Most discounted item,Footwear)}) are mined even if, for example, the support value of {(Most discounted item,Footwear)} is equal to the one of its descendant {(Most discounted item,Hiking boots)}. Let us consider now generalized itemset {(Most discounted item,Outerwear)}. It covers both the knowledge associated with its infrequent descendants {(Most discounted item,Ski pant)} and {(Most discounted item,Gloves)} and the knowledge represented by its frequent descendant {(Most discounted item,Jackets)}. Hence, evaluating the interestingness of {(Most discounted item,Outerwear)} is a challenging task. On the one hand, the pattern is interesting because it represents some knowledge that is not covered by any of its frequent descendants. However, on the other hand, the pattern expressiveness with respect to its descendants is rather limited, because the dataset records that contain jackets are already represented by the low-level itemset ({(Most discounted item,Jackets)}).

To select the high-level patterns that retain a significant degree of novelty with respect to their frequent descendants, we propose a novel type of high-level itemset, namely the EGI. While a traditional generalized itemset represents all of its descendant itemsets, both the frequent and the infrequent ones, each EGI represents, at a higher level of abstraction, the information represented by its subset of infrequent descendant itemsets. Hence, as thoroughly discussed in the following, EGIs are more expressive than traditional generalized itemsets because they consider, from a high-level viewpoint, only the rare knowledge that remains hidden at lower granularity levels.

EGI s are represented in the form $X \wr S$, where X is a generalized itemset and S is a set of (generalized) itemsets that contains only frequent descendants of X.[1] $X \wr S$ represents all the X's descendants, except for those contained in S, and covers all the records that are matched by X except for those already covered by any itemset in S. For example, {(*Most discounted item*,*Outerwear*)} $\wr$ {{(*Most discounted item*,*Jackets*)}} represents all the descendants of {(Most discounted item,Outerwear)}, except for {(Most discounted item,Jackets)}. Its support value is equal to 2, because it covers the same dataset records of {Outerwear} except for those that are covered by {(Most discounted item,Jacket)}. Symbol $\wr$, which is used to separate the X part from the S one, roughly recalls the symbol of complement between two sets ($\setminus$), because the knowledge covered by S is "excluded" from that covered by X. To enhance the readability and usability of the mined set, we consider a worthwhile EGI subset, i.e., the Max-EGIs. Max-EGIs are EGIs for which the S term has maximal length, i.e., the S term contains *all* the frequent descendants of X. Each Max-EGI compactly represents all the infrequent descendants of a generalized itemset X. Since the number of frequent descendants of X ($|S|$), is typically less than the number of infrequent descendants of X, the pattern $X \wr S$ is a compact yet expressive representation of the infrequent knowledge covered by X. By convenient abuse of set theory notation, the "complement" between a generalized itemset X and its frequent descendant set S represents the (potentially large) set of infrequent descendants of X. For example, {(*Most discounted item*,*Outerwear*)} $\wr$ {{(*Most discounted item*,*Jackets*)}} represents itemsets {(Most discounted item,Ski pant)} and {(Most discounted item,Gloves)}. Since they are individually infrequent but collectively frequent in the analyzed data, the expert could deem the above pattern to be interesting for advanced analysis. Table 2(b)

[1] For the sake of simplicity, in this preliminary example we neglect the case in which the S part may also contain other EGI s, which will be discussed in the following sections.

Download English Version:

https://daneshyari.com/en/article/393670

Download Persian Version:

https://daneshyari.com/article/393670

Daneshyari.com